Learning from Categorical Attribute Relationships for Positive-Unlabeled Classification

Learning from Categorical Attribute Relationships for Positive-Unlabeled Classification

Dino Ienco[1,2] and Ruggero G. Pensa[3]

[1] IRSTEA, UMR TETIS, Montpellier, France,
dino.ienco@irstea.fr
[2] LIRMM, Montpellier, France
[3] University of Torino, Dept. of Computer Science, Turin, Italy
ruggero.pensa@unito.it

Abstract. In common binary classification scenarios, the presence of both positive and negative examples in training data is needed to build an efficient classifier. Unfortunately, in many domains, this requirement is not satisfied and only one class of examples is available. To cope with this setting, classification algorithms have been introduced that learn from Positive and Unlabeled (PU) data. Originally, these approaches were exploited in the context of document classification. Only few works address the PU problem for categorical dataset. Nevertheless, the available algorithms are mainly based on Naive Bayes classifiers. In this work we present a new distance based PU learning approach for categorical data: *Pulce*. Our framework takes advantage of the intrinsic relationships between attribute values and exceeds the independence assumption made by Naive Bayes. *Pulce*, in fact, leverages on the statistical properties of the data to learn a distance metric employed during the classification task. We extensively validate our approach over real world datasets and demonstrate that our strategy obtains statistically significant improvements w.r.t. state-of-the-art competitors.

Keywords: metric learning, partially supervised learning, categorical data

1 Introduction

In common binary classification tasks, learning algorithms assume the presence of both positive and negative examples. Sometimes this is a strong requirement that does not fit real application scenarios. In fact, the process of labeling data is a money- and time-consuming activity that needs high-level domain expertise. In some cases this operation is quick, but usually, definining reliable labels for each data example is a hard task. In the worst case, extracting examples from one or more classes is simply impossible [3]. As a consequence, only a small portion of a so-constituted training set is labeled. As a practical example of this phenomenon, let us consider a company that aims at creating an archive of researchers' home

pages, using web-crawling techniques. Once downloaded, a web page should be classified to decide whether it is a researcher's home page or not. In such a context, the concept of positive example is well defined (the researcher's home page) while the idea of negative example is not well-established [12] because no real characterization of what is not a home page is supplied. The same problem occurs when trying to classify biological/medical data. Usually a biologist (or a doctor) can comfortably supply positive evidences of what she wants to identify but she is not able to provide negative examples. A known example of this scenario is the classification of vascular lesions starting from medical images [15], where labeling vascular lesions accurately could take more than one year, while it is relatively easy to recognize healthy individuals. In these scenarios, defining a method to exploit both positive and unlabeled examples could save precious material and human resources and the expert may focus her effort to only define what is good, skipping the ungrateful task of recognizing what is not good.

To deal with this setting, the Positive Unlabeled (PU) learning task has been introduced [4]. Roughly speaking, PU learning is a binary classification task where no negative examples are available. Most research works in this area are devoted to the classification of unstructured datasets such as documents represented by bag-of-words, but similar scenarios may occur with categorical data as well. Imagine, for instance, a dataset representing census records on a population. An analyst can comfortably provide reliable positive examples of a targeted class of people (e.g., unmarried young professional interested in adventure sports), but identifying plausible counterexamples is not as easy. However, very few PU learning approaches are designed to work on attribute-relation data (such as categorical datasets). Unfortunately the techniques proposed in text classification are not directly applicable to the context of attribute-relation datasets. These approaches, in fact, employ metrics, such as the cosine distance, that are not well suited for categorical data. Cosine distance, in fact, considers any mismatch of a binarized categorical variable in the same way. However, for a categorical variable such as "age class", value "child" is more similar to "teenager" than to "adult" or "senior". For categorical data, in addition, there is no standard definition of distance [1]. This limitation makes it impossible to apply works on document classification to categorical data directly. The few works that deal with PU learning in attribute-relation domains are principally based on Naive Bayes classifiers. The major limitation of this kind of approaches is that algorithms based on Naive Bayes assume that attributes are mutually independent. To the best of our knowledge, no effort was devoted to the implementation of other models or the extension of previously defined models from document analysis.

Contribution In this paper we introduce a new distance-based algorithm, named *Positive Unlabeled Learning for Categorical datasEt* (*Pulce*). Our work aims at filling the gap between the recent and well-established advances in document classification and the preliminary status of works existing for attribute-

relation data. In particular, we address the problem of classifying data described by categorical attributes, which also includes the case of discretized numerical attributes, leading to a general framework for attribute-relation data. The core part of our approach is an original distance-based classification method which employs a distance metric learnt directly from data thanks to a technique recently presented by Ienco *et al.* [9]. Originally, this technique was designed to exploit attribute dependencies in an unsupervised (clustering) scenario, and allows to quantify the distance between any pair of values of the same categorical attribute X_i by the way in which the values of the other attributes X_j are distributed in the dataset examples: if they are similarly distributed in the groups of samples in correspondence of the distinct values of X_i a low value of distance is obtained. Our PU learning approach uses this metric to train two discriminative models: one for the positive class, the other for the negative one. These two models take intrinsically into account the existing attribute relationships, thus overcoming the major limitation given by the independence assumption explicitly made by Naive Bayes-based methods. We provide the empirical evidence of this property, showing that our method outperforms state-of-the-art competitors and assessing the statistical significance of the results.

Related work Positive Unlabeled learning was originally studied by De Comité *et al.* [4] who achieved the first theoretical results. The authors showed that under the PAC (Probably Approximately Correct) learning model, the k-DNF (k-Disjunction Normal Form) approach is able to learn from positive and unlabeled examples. Following these preliminaries results, PU learning was first applied to text document classification [12].

Other approaches dealing with PU classifiers in the context of text classification have been presented in more recent years [6, 11, 14]. Elkan *et al.* [6] introduce a SVM-based classifier method that assign weights to the examples belonging to the unlabeled set. Xiao *et al.* [11] combine two techniques borrowed from information retrieval (Rocchio and Spy-EM) to extract a set of reliable negative examples. Finally, Zhou *et al.* [14] modify the standard Topic-Sensitive probabilistic Latent Semantic Analysis (pLSA) approach to perform classification with a small set of positive labeled examples.

Calvo *et al.* [2] first attempted to deal with the PU learning setting in attribute-relation datasets. Their paper introduces four methods based on Naive Bayes for categorical data. In particular the authors modify classic and Tree Augmented Naive Bayes [7] approaches to work with positive and unlabeled examples. They supply two ways to estimate the a priori probability of the negative class: the first one takes into consideration the whole set of unlabeled examples to derive this probability, while the second one considers a Beta distribution to model the uncertainty. These methods are substantially limited by two aspects: the strong (and often wrong) assumption of attribute independency adopted by Naive Bayes and the use of the whole set of unlabeled examples to estimate a model for the negative class. This work is extended by He *et al.* [8] to deal with uncertainty data.

In conclusion, even though much work has been devoted to document classification, and some effort exists for specific kinds of applications, very few researches address the problem of building reliable classifiers over positive and unlabeled examples in attribute-relation data. To the best of our knowledge, our work is the first one trying to cope with PU learning outside the document classification domain without any strong (and often wrong) attribute independency assumption.

Paper organization The remainder of this paper is organized as follows: The problem formulation, a brief overview of the distance learning algorithm, and the full description of the proposed method are supplied in Section 2. In Section 3 we provide our empirical study and analyze its statistical significance. Finally, Section 4 concludes.

2 A distance-based method for categorical data

In this section we introduce PU learning and describe *Pulce*, a new distance based PU learning schema for categorical data. After some general definitions, we briefly describe the distance learning framework we adopt in our approach. Then, we provide the technical details of our distance-based PU learning algorithm.

We consider a dataset $D = \{P \cup U\}$ composed by a set P of positive examples and a set U of unlabeled examples all described by a set $F = \{X_1, X_2, \ldots, X_m\}$ of m categorical attributes. The task of learning from both positive and unlabeled examples consists in exploiting both labeled P and unlabeled U examples to learn a model allowing the assignment of a label to new, previously unseen, examples. The general process is performed in two steps:

1. detect a reliable set of negative examples $RN \subseteq U$;
2. build a classifier over $\{P \cup RN\}$.

The key intuition behind our approach is that, if we learn a distance based on positive examples only, negative examples will be differently distributed w.r.t. this metric. In other terms, negative examples would not fit the learnt distance model, and they will be easily detected and labeled as reliable negative. Following this preamble, we employ the distance learning framework for categorical data presented by Ienco *et al.* [9] to learn a distance model for the attributes in F on the sole set of positive example P. This distance model is used to weight each unlabeled example in U. A cut-off threshold is then automatically computed, and a set RN of reliable negative examples is supplied to a distance-based classifier. In particular, we adopt a modified version of k-NN that projects each test example into each class space using the corresponding metric model and evaluates the k nearest neighbors around it. The test example is assigned the class that minimizes the sum of distances w.r.t. its k nearest neighbors.

In the following, we will first recall briefly the distance learning method adopted in our framework. Then, we describe the ranking strategy used to identify a reliable set of negative examples. Finally, we introduce the classification algorithm.

ID	Age	Gender	Profession	Product	Sale dep.
1	young	M	student	mobile	suburbia
2	senior	F	retired	mobile	suburbia
3	senior	M	retired	mobile	suburbia
4	young	M	student	smartphone	suburbia
5	senior	F	businessman	smartphone	center
6	adult	M	unemployed	smartphone	suburbia
7	adult	F	businessman	tablet	center
8	young	M	student	tablet	center
9	senior	F	retired	tablet	center
10	senior	M	retired	tablet	center

(a) Sales table

	mobile	smartphone	tablet
student	1	1	1
unemployed	0	1	0
businessman	0	1	1
retired	2	0	2

(b) Product-Profession contingency table

	mobile	smartphone	tablet
center	3	1	0
suburbia	0	2	4

(c) Product-Sales dep. contingency table

Fig. 1. *Sales*: a sample dataset with categorical attributes (a) and two related contingency tables (b and c).

2.1 Computing the distance model

Here we briefly summarize *DILCA* (DIstance Learning for Categorical Attributes), a framework for computing distances between any pair of values of a categorical attribute. *DILCA* was introduced by Ienco *et al.* [9], but was limited to a clustering scenario.

To illustrate this framework, we consider the dataset described in figure 1(a), representing the set *Sales*. It has five categorical attributes: *Age*{*young, adult, senior*}, *Gender*{*M,F*}, *Profession*{*student,unemployed,businessman,retired*}, *Product*{*mobile,smartphone,tablet*} and *Sales department*{*center,suburbia*}. The contingency tables in Figure 1(b) and Figure 1(c) show how the values of attribute *Product* are distributed w.r.t. the two attributes *Profession* and *Sales department*. From Figure 1(c), we observe that *Product=tablet* occurs only with *Sales dep.=suburbia* and *Product=mobile* occurs only with *Sales dep.=center*. Conversely, *Product=smartphone* is satisfied both when *Sales dep.=center* and *Sales dep.=suburbia*. From this distribution of data, we infer that, in this particular context, *tablet* is more similar to *smartphone* than to *mobile* because the probability to observe a sale in the same department is closer. However, if we take into account the co-occurrences of *Product* values and *Profession* values (Figure 1(b)), we may notice that *Product=mobile* and *Product=tablet* are closer to each-other rather than to *Product=smartphone*, since they are bought by the same professional categories of customers to a similar extent.

This example shows that the distribution of the co-occurrence table may help to define a distance between values of a categorical attribute, but also that the context matters. Let us now consider the set $F = \{X_1, X_2, \ldots, X_m\}$ of m categorical attributes and dataset D in which the instances are defined over F. We denote by Y the target attribute, which is a specific attribute in F that

constitutes the target of the method, that is, on whose values we need to compute the distances. *DILCA* allows to compute a context-based distance between any pair of values (y_i, y_j) of the target attribute Y on the basis of the similarity between the probability distributions of y_i and y_j given the context attributes, called $\mathcal{C}(Y) \subseteq F \setminus Y$. For each context attribute X_i it computes the conditional probability for both the values y_i and y_j given the values $x_k \in X_i$ and then it applies the Euclidean distance. The Euclidean distance is normalized by the total number of considered values:

$$d(y_i, y_j) = \sqrt{\frac{\sum_{X \in \mathcal{C}(Y)} \sum_{x_k \in X} (P(y_i|x_k) - P(y_j|x_k))^2}{\sum_{X \in \mathcal{C}(Y)} |X|}} \quad (1)$$

The selection of a good context is not trivial, particularly when data are high-dimensional. To select a relevant and non redundant set of features w.r.t. a given one, The authors [9] propose to adopt *FCBF*, a feature-selection approach originally presented by Yu and Liu [13]. Here we use exactly the same strategy, which is based on the *relevance* and the *redundancy* criteria between attributes. To evaluate the correlation for both *relevance* and *redundancy* they employ the *Symmetric Uncertainty* measure (*SU*). *SU* is a normalized version of the *Information Gain* [10] and it ranges between 0 and 1. Given two variables X and Y: 1 indicates that knowledge of the value of either Y or X completely predicts the value of the other variable; 0 indicates that Y and X are independent. During the step of context selection, a set of context attributes $\mathcal{C}(Y)$ for a given target attribute Y is selected. Informally, these attributes $X_i \in \mathcal{C}(Y)$ should have a high value of $SU(Y, X_i)$ and are not redundant among them. $SU_Y(X_i)$ denotes the Symmetric Uncertainty of X_i for the target Y. *DILCA* first produces a ranking of the attribute X_i in descending order w.r.t. $SU(Y, X_i)$. This operation implements the *relevance* step. Starting from the ranking, it compares each pairs of ranked attributes X_i and X_j. One of them is considered redundant if the Symmetrical Uncertainty that links them is higher than the Symmetrical Uncertainty that links each of them to the target. In particular, X_j is removed if X_i is in higher position of the ranking w.r.t. X_j ($SU_X(X_j) < SU_Y(X_i)$) and the Symmetric Uncertainty that links them is higher than the SU that links each of them to the target ($SU_{X_j}(X_i) > SU_Y(X_i)$ and $SU_{X_j}(X_i) > SU_Y(X_j)$). This second part of the approach implements the *redundancy* step. The results of the whole procedure is the set of attributes $\mathcal{C}(Y)$.

At the end of the process, *DILCA* returns a distance model $\mathcal{M} = \{M_{X_i} \mid i = 1, \ldots, m\}$, where each M_{X_i} is the matrix containing the distances between any pair of values of attribute X_i, computed using Eq. 1.

2.2 Detecting reliable negative examples

Here we present our solution to the problem of extracting a set of reliable negative examples from U. The whole procedure is sketched in Algorithm 1. As first step, we learn a distance model $\mathcal{M}_P$, using *DILCA* on P (see Section 2.1). $\mathcal{M}_P$ summarizes the relationships between attributes in P in such a way that

Algorithm 1: $Pulce(P,U)$

$\mathcal{M}_P \leftarrow DILCA(P)$;
$\tau \leftarrow \frac{2}{|P|(|P|-1)} \sum_{i=1}^{|P|-1} \sum_{j=i+1}^{|P|} dist(\mathcal{M}_P, p_i, p_j)$;
$RN \leftarrow \{\emptyset\}$;
forall the $u \in U$ **do**
  if $(score(u, P, \mathcal{M}_P) > \tau)$ **then**
    $RN \leftarrow RN \cup u$;
  end
end
$\mathcal{M}_{RN} \leftarrow DILCA(RN)$;
return $\mathcal{M}_P$, $\mathcal{M}_{RN}$, RN;

new examples drawn from the same distribution will be closer to P than new examples drawn from a different distribution.

Using the model $\mathcal{M}_P$, for each example $u \in U$, we compute a score based on the average distance between u and all examples $p \in P$. The score is computed as follows:

$$score(u, P, \mathcal{M}_P) = \frac{\sum_{p \in P} dist(\mathcal{M}_P, u, p)}{|P|} \tag{2}$$

where the function $dist(\mathcal{M}_P, u, p)$ could be any distance function that uses only the distance between two values of the same attribute. In our case we use the Euclidean distance:

$$dist(\mathcal{M}_P, u, p) = \sqrt{\sum_{M_{X_i} \in \mathcal{M}_P} M_{X_i}(u[X_i], p[X_i])^2} \tag{3}$$

where $u[X_i]$ and $p[X_i]$ are the values the attribute X_i takes in examples u and p respectively, so that $M_{X_i}(u[X_i], p[X_i])$ is the distance between values $u[X_i]$ and $p[X_i]$ of attribute X_i.

Multiple different choices can be adopted for the selection of a reliable set of negative examples given this score. A first possibility is to rank all examples $u \in U$ in decreasing order of score. Hence, examples from the negative class are likely to be on top of the ranking and the user may decide to label the first n examples as reliable negative. Instead, we provide a strategy to select a reliable set RN of negative examples automatically: we mark as reliable negative all examples $u \in U$ such that the $score(u, P, \mathcal{M}_P)$ is greater than a threshold τ, i.e., $RN = \{u \in U \ \ s.t. \ \ score(u, P, \mathcal{M}_P) > \tau\}$. The problem now is how to tune correctly the value of τ in order to detect reliable negative examples. Even though sophisticated strategies could be adopted, here we consider a simple solution: we employ the mean of all distances within the set P:

$$\tau = \frac{2}{|P|(|P|-1)} \sum_{i=1}^{|P|-1} \sum_{j=i+1}^{|P|} dist(\mathcal{M}_P, p_i, p_j) \tag{4}$$

where $|P|$ is the cardinality of the set P of positive examples.

2.3 Classifying positive and reliable negative examples

We now dispose of the set P of positive examples and the set RN of reliable negative examples, and we are able to build a discriminative model to recognize and label new unseen examples. To perform our classification task we use a revised k-NN (k nearest neighbors) approach. In particular, the major difference with standard k nearest neighbors approaches consists in the adoption of two different distances, one for the positive class and one for the negative class. Each distance learnt by *DILCA* constitutes a way to summarize the attribute dependencies within each class. This enables *Pulce* to build a specific model for each class. Concerning the positive class, we use distance model $\mathcal{M}_P$. For the negative class, we learn a distance model $\mathcal{M}_{RN}$ by applying *DILCA* on the set RN of reliable negative examples. The key intuition behind our classification method is the following: if a new, unseen example t comes from a specific class, the corresponding distance model should produce small distances with other examples from its class w.r.t. other distance models learnt from other classes. The classifier then considers the k examples from each class that are closest to t. Finally, for each class, it sums the distances between the unseen example t and its k nearest neighbors and assign it the class that minimizes this value. The advantage of learning two distance models is now clear. A classifier based on a unique model requires the definition of a threshold (or other more sophisticated strategies) to decide whether an example can be considered positive or negative. The use of a distance model for each class makes this complex step unnecessary. Notice that, as *DILCA* provides values that are bounded between 0 and 1, the two distances are comparable.

We formalize our nearest neighbors approach as follows. Given an unseen example t, we call $NN_P(t) = \{nn_1^P(t), \ldots, nn_k^P(t)\}$ the set of k nearest neighbors of t in P under the distance model $\mathcal{M}_P$, and $NN_{RN}(t) = \{nn_1^{RN}(t), \ldots, nn_k^{RN}(t)\}$ the set of k nearest neighbors of t in RN under the distance model $\mathcal{M}_{RN}$. Then, the class of t is given by:

$$class(t) = \underset{c \in \{P, RN\}}{\operatorname{argmin}} \sum_{i}^{k} dist(\mathcal{M}_c, nn_i^c(t), t) \tag{5}$$

Notice that k is the only parameter of the whole PU learning approach, as in many instance-based classifiers.

3 Experimental results

In this section we provide an exhaustive set of experiments to show the effectiveness of our PU learning approach in categorical data. The experiments are performed over 39 samples derived from 13 datasets, publicly available on the UCI machine learning repository[1]. For each dataset we produce three different samples that differ from each other in the number of examples labeled as

[1] http://archive.ics.uci.edu/ml/

positive, respectively 30%, 40%, and 50% of the positive class. The remaining positive examples plus all the negative examples are considered as unlabeled instances. We assume the majority class as positive, the other one as negative. If the dataset does not describe a binary classification problem we select the two biggest classes (in the number of instances) to reduce the problem to a binary classification task. Finally, as further pre-processing, all numerical attributes are discretized into 10 bins with equal width. The details on all the 39 samples are presented in Table 1. To evaluate the results of the different PU classifiers we use the F-Measure as performance indicator.

3.1 Comparative results

We choose to compare *Pulce* with the four approaches based on Naive Bayes proposed by Calvo *et al.* [2]: Positive Naive Bayes *PNB*, Average Positive Naive Bayes *APNB* (both based on classic Naive Bayes), Positive TAN (*PTAN*) and Average Positive TAN (*APTAN*), two variants of the tree augmented Naive Bayes model [7]. The difference between *PNB* (resp. *PTAN*) and *APNB* (resp. *APTAN*) lies in the way the *a priori* probability for the negative class is estimated. For *PNB* and *PTAN* this probability is derived directly from the unlabeled set of examples while for *APNB* and *APTAN* the uncertainty is modelled by a Beta distribution. For these four approaches, we use the standard parameter settings, as suggested by Calvo *et al.* [2]. For *Pulce*, instead, we use $k = \{1, 3, 7, 11, 13\}$.

The results are reported in Table 1 (because of space limitations, only detailed table for $k = 7$ is presented here), where we can observe the performances of *Pulce* in comparison with those of the competitors. In general, the first remarkable result is that *Pulce* outperforms the other methods both in terms of average F-Measure and in terms of number of wins, independently of the value of k. In detail, it wins 26 times (for $k = 1$), 25 times ($k = 3$), 29 times ($k = 7$), 25 ($k = 11$) and 28 times (when $k = 13$). The best competitor (*APNB*) wins only 10 times, no matter which value of k we consider. Furthermore, *Pulce*'s average F-Measure (0.72, for $k = \{1, 3\}$ and 0.73 for $k = \{7, 11, 13\}$) is sensibly higher than competitors' one: *APNB* and *PNB* do not go above 0.67. Notice also that, when *Pulce* is not the best PU method, its results are in line with the ones achieved by the competitors.

We may also observe that in *chess*, *dermatology*, *mushroom*, *nursery* and *vote* the improvements w.r.t. the F-Measure evaluated for the competitors are clearly visible. This result is somehow expected and it is due to the fact that these datasets are dense and correlated. *Pulce* exploits the dependencies among attributes and overcomes the limitation of the Naive Bayes model, which is funded on the independence assumption. In general, this assumption is wrong, especially in the dataset listed above. Results on *audiology* deserve some comments as well. This dataset is relatively high-dimensional. Few data instances (226) are described by an high number of attributes (69). This is also a limitation for Naive Bayes approaches, but not for *Pulce*, that is able to exploit attribute dependency and, in this case, outperforms all other competitors by far.

Dataset	N. Attr.	% pos.	N. Pos.	N. Unlab.	*PNB*	*PTAN*	*APNB*	*APTAN*	*Pulce*
audiology	69	30%	15	79	0.68	0.66	0.70	0.66	**0.74**
		40%	20	74	0.75	0.71	0.74	0.66	**0.85**
		50%	26	68	0.80	0.78	0.80	0.71	**0.90**
breast-cancer	9	30%	54	203	0.40	0.43	0.39	0.43	**0.53**
		40%	72	185	0.42	0.43	0.40	**0.45**	0.44
		50%	91	166	0.42	**0.44**	0.41	**0.44**	**0.44**
chess	36	30%	451	2425	0.58	0.59	0.64	0.64	**0.70**
		40%	601	2275	0.58	0.60	0.64	0.64	**0.69**
		50%	751	2125	0.58	0.60	0.64	0.64	**0.66**
connect4	9	30%	169	693	0.47	0.48	0.47	0.39	**0.54**
		40%	225	637	0.48	0.51	0.47	0.42	**0.57**
		50%	282	580	**0.49**	**0.49**	0.48	0.43	0.41
dermatology	34	30%	30	136	0.57	0.57	0.57	0.56	**0.99**
		40%	40	126	0.57	0.57	0.58	0.57	**0.99**
		50%	50	116	0.59	0.57	0.60	0.58	**0.99**
heart-c	13	30%	45	228	**0.73**	0.63	0.70	0.64	0.72
		40%	60	213	0.77	0.63	**0.78**	0.70	0.75
		50%	75	198	**0.77**	0.63	**0.77**	0.68	**0.77**
hepatitis	19	30%	9	130	**0.87**	0.85	**0.87**	0.86	**0.87**
		40%	12	127	**0.88**	0.85	**0.88**	0.85	**0.88**
		50%	15	124	**0.88**	0.86	**0.88**	0.85	**0.88**
lymph	18	30%	17	111	0.84	0.79	**0.85**	0.84	0.84
		40%	22	106	**0.84**	0.79	0.83	0.81	0.83
		50%	28	100	0.86	0.81	**0.87**	0.82	0.80
mushroom	22	30%	1136	6176	0.72	0.68	0.67	0.67	**0.74**
		40%	1515	5797	0.75	0.73	0.67	0.68	**0.76**
		50%	1894	5418	0.75	0.74	0.68	0.69	**0.82**
nursery	8	30%	1166	6562	0.65	0.56	0.65	0.50	**0.74**
		40%	1555	6173	0.69	0.61	0.69	0.56	**0.77**
		50%	1944	5784	0.69	0.74	0.70	0.44	**0.81**
pima	8	30%	135	556	0.49	0.50	0.50	0.50	**0.53**
		40%	180	511	0.49	0.50	0.50	0.51	**0.55**
		50%	225	466	0.49	0.50	0.51	0.52	**0.52**
soybean	35	30%	25	140	0.81	0.80	**0.86**	0.81	0.73
		40%	33	132	**0.86**	0.84	**0.86**	0.83	0.76
		50%	42	123	**0.92**	0.88	**0.92**	0.86	0.83
vote	16	30%	72	319	0.62	0.56	0.62	0.55	**0.68**
		40%	96	295	0.71	0.58	0.71	0.54	**0.80**
		50%	120	271	0.77	0.61	0.77	0.56	**0.82**
N. of wins					9	2	10	2	**29**
Avg. F-Meas.					0.67	0.64	0.67	0.62	**0.73**
Avg. ranking					2.884	3.679	2.756	3.833	**1.846**

Table 1. F-Measure results over the 39 samples with $k = 7$ ($\chi_F^2 = 39.7432$)

Finally, let us make some comments concerning the number of positive examples involved in the learning step. The accuracy of this kind of approaches should improve as the number of available positive examples grows, according to the theory that, with a large enough set of positive examples the performance of PU classifiers could be the same of standard binary classifiers learnt over both positive and negative examples [4]. From Table 1 it turns out that this is not always the case in our experiments. Except for *connect4* and *chess* which are related to strategy games, other factors that should be taken into account are related to under/overfitting phenomena. Our approach has two learning phases: the distance learning step and the k-NN step. Each of these steps suffers from typical (and sometimes unpredictable) classification biases. In some cases, too many positive examples may bias the classification task towards a major accuracy for the positive class. In some other cases, the problem is inverted. However, it can be noticed that when the number of positive examples is low (30%) our approach wins 9 times over 13 for $k = \{1, 3, 11\}$ and 10 times over 13 for $k = \{7, 13\}$, for a win-ratio which is always higher than the overall win-ratio.

3.2 Statistical significance of the results

To asses the statistical quality of our approach we use the Friedman statistics and the Nemenyi test [5]. These techniques are usually employed to deal with the problem of evaluating the statistical relevance of results of different classifiers over multiple datasets. We compare *Pulce* with all the competitors (*PNB*, *APNB*, *PTAN*, *APTAN* over 13 datasets with 3 different percentage of labeled positive examples, for a total of 39 datasets. In this statistic test the null hypothesis is that all the methods obtains similar performances, i.e., the χ^2_F value is similar to the critical value for the chi-square distribution with 4 degrees of freedom. At significance level of $\alpha = 0.001$, the critical values of the chi-square is equal to 18.467. In our test we obtain values from 27.1488 to 39.7432 for the χ^2_F statistics (see captions of Table 1, hence the null hypothesis of the Friedman test is comfortably rejected and we can now proceed with the post-hoc Nemenyi test. According to this test, the performance of two classifiers is significantly different if the corresponding average ranks differ by at least the critical difference. Then, we compare the five methods at the critical value of $q_{0.1} = 2.459$. The ranking table is shown on bottom of Table 1. The critical difference is $CD = 0.8805$ (at significance level $\alpha = 0.1$). We observe that *Pulce* brings a statistically significant improvement w.r.t. all the four competitors, with an average rank of 1.846.

4 Conclusion

We have introduced *Pulce*, a distance based approach for Positive Unlabeled learning (PU) in categorical data. Unlike the few existing approaches based on Naive Bayes, it takes into account data dependencies and learns a distance model from attribute relationships to train a k-NN-like classifier. Statistically significant classification results on real world datasets have validated our strategy, also

comparing to state-of-the-art approaches, Finally, the sensibility analysis on the only required parameter and the study on variations of the proposed threshold for selecting a good example-set for the negative class and have shown that *Pulce* is stable and robust.

References

1. Boriah, S., Chandola, V., Kumar, V.: Similarity measures for categorical data: A comparative evaluation. In: Proceedings of SDM 2008, Atlanta, GA, USA, SIAM, Philadelphia, PA, USA (24-26 April 2008) 243–254
2. Calvo, B., Larrañaga, P., Lozano, J.A.: Learning bayesian classifiers from positive and unlabeled examples. Pattern Recognition Letters **28**(16) (2007) 2375–2384
3. Calvo, B., López-Bigas, N., Furney, S.J., Larrañaga, P., Lozano, J.A.: A partially supervised classification approach to dominant and recessive human disease gene prediction. Computer Methods and Programs in Biomedicine **85**(3) (2007) 229–237
4. Comité, F.D., Denis, F., Gilleron, R., Letouzey, F.: Positive and unlabeled examples help learning. In: Proceedings of ALT 1999, Tokyo, Japan, Springer, Berlin (6-8 December 1999) 219–230
5. Demsar, J.: Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research **7** (2006) 1–30
6. Elkan, C., Noto, K.: Learning classifiers from only positive and unlabeled data. In: Proceedings of KDD 2008, Las Vegas, Nevada, ACM, New York, NY, USA (24-27 August 2008) 213–220
7. Friedman, N., Geiger, D., Goldszmidt, M.: Bayesian network classifiers. Machine Learning **29**(2-3) (1997) 131–163
8. He, J., Zhang, Y., Li, X., Wang, Y.: Naive bayes classifier for positive unlabeled learning with uncertainty. In: Proceedings of SDM 2010, Columbus, Ohio, SIAM, Philadelphia, PA, USA (29 April - 1 May 2010) 361–372
9. Ienco, D., Pensa, R.G., Meo, R.: From context to distance: Learning dissimilarity for categorical data clustering. Trans. in Knowledge Discovery from Data **6**(1) (mar 2012) 1:1–1:25
10. Quinlan, R.J.: C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann, Burlington, MA, USA (1993)
11. Xiao, Y., Liu, B., Yin, J., Cao, L., Zhang, C., Hao, Z.: Similarity-based approach for positive and unlabeled learning. In: Proceedings of IJCAI 2011, Barcelona, Spain, AAAI, Palo Alto, CA, USA (16-22 July 2011) 1577–1582
12. Yu, H., Han, J., Chang, K.C.C.: Pebl: positive example based learning for web page classification using svm. In: Proceedings of KDD 2002, Edmonton, Alberta, Canada, ACM, New York, NY, USA (23-26 July 2002) 239–248
13. Yu, L., Liu, H.: Feature selection for high-dimensional data: A fast correlation-based filter solution. In: Proceedings of ICML 2003, Washington, DC, USA, AAAI, Palo Alto, CA, USA (21-24 August 2003) 856–863
14. Zhou, K., Xue, G.R., Yang, Q., Yu, Y.: Learning with positive and unlabeled examples using topic-sensitive plsa. IEEE Trans. Knowl. Data Eng. **22**(1) (2010) 46–58
15. Zuluaga, M.A., Hush, D., Leyton, E.J.F.D., Hoyos, M.H., Orkisz, M.: Learning from only positive and unlabeled data to detect lesions in vascular ct images. In: Proceedings of MICCAI 2011, Toronto, Canada, Springer, Berlin (18-22 September 2011) 9–16